Tapping into the Collective Creativity of the Crowd: The Effectiveness of Key Incentives in Fostering Creative Crowdsourcing

Ioana Literat
Teachers College, Columbia University
literat@tc.columbia.edu

Abstract

To better understand the conditions that most effectively stimulate creative participation online, a crowdsourcing project was implemented on Amazon's Mechanical Turk, collecting 4200 written and visual submissions from online participants. An experimental research design tested the impact of specific incentive structures (i.e. financial rewards, bonuses, specification of project purpose, attribution of authorship credit) on the outcomes of creative participation (quantity of submissions, quality of submissions, time spent on task). The study found that extrinsic rewards (i.e. higher pay and bonuses) are effective in encouraging participants to accept the creative task, whereas the strategies that boost the creativity of the submissions are: offering a bonus, mentioning a charitable purpose, and giving contributors authorship credit. These findings help illuminate the factors that have the greatest impact on the quality and quantity of online creative participation, thus making a vital contribution to our understanding of digital creativity.

1. Introduction

Defined by its coiner, Jeff Howe, as "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call" [1], crowdsourcing began as a corporate strategy of engaging customers in the generation of creative content, innovation and brand development. Recent years have seen a proliferation of creative projects that are based on open public participation and take place entirely online [2]. As the reliance on crowdsourced creativity becomes an increasingly common practice in both commercial and non-commercial contexts, there is a need to achieve a better understanding of the factors that drive creative participation online, and the possible strategies that might increase the effectiveness of creative crowdsourcing initiatives.

Thus, the principal goal of the present study was to investigate, through an experimental design, the conditions that most effectively foster creative participation in online spaces, by looking at the incentive structures that shape creative collaboration in online environments. Specifically, participants on the crowdsourcing platform Mechanical Turk were asked to contribute to the writing and illustration of a children's book about digital culture - a children's book about the Internet, by the Internet. A total of 4200 creative contributions were collected, out of which half were written submissions and half were illustrations. The aim was to assess the impact of four key factors (amount of reward; possibility of bonus payment; stated purpose of the project; attribution of authorship credit) on the quantity and quality of the resulting submissions, and the time spent by participants on task.

By determining the impact that various incentives have on the quality and quantity of creative participation online, this work makes a vital contribution to our understanding of participant motivations in crowdsourcing. The outcome of this investigation will thus be useful not only to researchers, but also to organizations who aim to rely on online participation for their creative projects; indeed, a recent report on the state of crowdsourcing found that 85% of the best global brands have used crowdsourcing in the past 10 years, and the popularity of the practice is continuing to rise [3]. Significantly, this research will also help gain a better understanding of participant motivations in the specific context of creative crowdsourcing applications, thus building a stronger bridge between crowdsourcing and creativity research.

2. Background

As crowdsourcing continues to gain prominence across a wide variety of fields, from product innovation [4] to social science research [5] to participatory art and culture [2], determining the

URI: http://hdl.handle.net/10125/41366
ISBN: 978-0-9981331-0-2

incentive structures that drive online participation has been important goal of recent crowdsourcing scholarship [6], [7], [8]. Having a better understanding of the strategies that work best to facilitate the success of crowdsourcing campaigns is particularly important because, as Simula notes, crowdsourcing initiatives encounter significant challenges in attracting and maintaining quality participation, and for every successful campaign, there are many failed ones as well [9].

On a basic level, individuals are motivated by intrinsic and extrinsic forces, although these are not mutually exclusive. *Intrinsic motivation* is defined as the "inherent tendency to seek out novelty and challenges, to extend and exercise one's capacities to explore, and to learn" [10], while e*xtrinsic motivation* means "doing something because it leads to a separable outcome" [11]. This tension is particularly relevant to creative crowdsourcing initiatives: research from creativity studies has shown that, in contexts of creative engagement, intrinsic incentives are most often stronger than extrinsic ones, and, furthermore, in some cases the latter can "crowd out" the former, decreasing performance and enjoyment [12], [13], [14]. Furthermore, in their study of worker performance on Mechanical Turk, Rogstadius and colleagues found that the quality of work increased when intrinsic motivation was stronger than extrinsic motivation [15].

Of particular importance is the impact of monetary incentives on participation. Research has found that financial rewards play a significant role in getting workers to accept tasks [8], but do not necessarily lead to an increase in the quality of work [16]. Another interesting finding has been that using bonus rewards for best ideas - as a tactic to increase the quantity and quality of submissions generated - does not always have the desired effect, and can sometimes create "a conflict of goals, as rewards can have perverse effects on the outcome" [17].

However, while financial incentives continue to be a significant motivating factor [8], studies have found that, depending on the context and the nature of participation, contributors are also motivated by incentives other than financial rewards, or a combination between financial and non-financial incentives [6], [18], [19], [20]. According to management researchers at MIT's Center for Collective Intelligence, the incentives that motivate crowdsourcing participants can be broadly classified as money, love, and glory, or a combination of the three [18]. Researching the community at Threadless.com, Brabham similarly found four principal motivators for participation: "the opportunity to make money, the opportunity to develop one's creative skills, the potential to take up freelance work, and the love of community" [6]. In their study of 12 open innovation platforms, Antikainen, and Väätäjä concluded that both monetary and non-monetary rewards can stimulate participation, and note the value of using a combination of both [19]. Shaw and colleagues reached a similar conclusion about the effectiveness of combining material and non-material incentives on Mechanical Turk [20].

In regards to the main non-financial incentives that drive participation in crowdsourcing, Kaufmann and colleagues identified three main types in their overview of the crowdsourcing literature: enjoyment-based motivations like having fun and passing time, community-based motions like social interaction and community identification, and social motivations like seeking social approval [21]. The significance of intrinsic motivations on Mechanical Turk is echoed by the findings of similar studies [7], [15], [22], [23].

A key incentive with a strong potential to shape participation in crowdsourcing projects is the purpose of the task. In a very interesting study of the relationship between task meaningfulness and motivation to participate on Mechanical Turk, Rogstadius and colleagues found that worker performance was more accurate when the task was framed as aiding a non-profit versus a corporate entity [15]. Chandler and Kapelner similarly found that US workers were more inclined to perform a task if they perceived it as socially meaningful (in this case, labeling tumor cells to aid a groundbreaking cancer treatment), but Indian workers were not influenced by the perceived social utility of the task [7].

While there is no previous empirical research on the role of authorship as incentive in crowdsourcing initiatives, authorship credit – or lack thereof – was identified as a significant source of conflict between requestors and participants in research on crowdsourcing in the artistic field [2]. Therefore, given this perceived significance and the lack of existing research on this particular dynamic, authorship credit was included as a variable here, in an effort to generate important findings in this regard.

3. Study Description

This study aims to contribute to this growing body of research by testing the impact of key incentive structures (i.e. financial rewards, bonuses, specification of project purpose, attribution of authorship credit) on the outcomes of creative participation (quantity of submissions, quality of submissions, and time spent on task).

Based on the literature surveyed above, it was hypothesized that the quantity (H1) and quality - i.e.

usefulness (H2) and novelty (H3) - of the submissions, as well as the time spent on task (H4) would increase when:

a. there is a potential bonus payment.

b. the stated purpose of the task is charity, and will decrease when the stated purpose is commercial.

c. there is the possibility of gaining credit for one's work, and will decrease when there is no attribution of authorship credit.

d. the amount of the financial reward is greater, and will decrease when the reward is lower.

3.1. Platform and Participants

Developed by Amazon in 2005, the Mechanical Turk website allows individual requesters to post "human intelligence tasks" ("HITs"), which online workers ("turkers") complete in exchange for a fee. Common HITs require workers to tag visual or written content, transcribe audio, or answer surveys. Generally, Mechanical Turk tasks are fast, easy and often repetitive; the monetary rewards for workers are low, most tasks being worth only a few cents. Due to its large user base and the low costs required to elicit participation, the platform has become a popular tool for researchers, especially in the social sciences.

According to Amazon, the Mechanical Turk platform has over 500,000 registered workers [24]; realistically, however, the number of active turkers regularly signing in and working on tasks must be much lower. Given the scope of this study, no demographic data was collected about the participants (although many chose to self-disclose information about their gender and occupation in the two open-ended questions in my HITs; see discussion section). Research on the general demographics of Mechanical Turk indicates that most workers are from the US (46.8%) and India (34%). In terms of gender, the majority of American workers are female, while in India the situation is reversed. When compared to the general population of Internet users, Mechanical Turk workers tend to be younger and better educated, while income levels are, roughly, similarly distributed [25].

3.2. Materials and Procedure

This study used an experimental design whereby 42 different tasks ("HITs", in the Mechanical Turk terminology) were posted on Mechanical Turk. Each HIT allowed for 100 responses ("assignments"); thus, a HIT was considered completed ("expired") when all 100 responses were submitted.

Two basic task templates were created: one for writing tasks and one for illustration tasks. Both tasks mentioned that the workers were participating in writing - or, respectively, illustrating - a children's book about a snail called Hashtag. The writing task provided an initial setup ("On his way home, Hashtag the Snail stumbled upon another snail's shell. He looked around, wondering who this mysterious shell belonged to, but there was no other snail in sight…") and asked participants to continue the story, exquisite-corpse-style, by providing the next sentence. It was specified that, in terms of the storyline, "the only requirement is that Hashtag must somehow use the Internet to accomplish his goal." The illustration task provided the same initial sentence and asked workers to draw and then upload an image that accurately represents that narrative situation. The HIT further specified all types of visual depiction were welcome: digital illustrations, drawings made by hand and then scanned or photographed, and even collages, ASCII art or any other way they would choose to depict Hashtag's story. The range of accepted media was deliberately broad, so as not to exclude workers who do not have experience with digital illustration software.

Building on these two basic task templates (i.e. writing and illustration), for each experimental condition I modified the following factors (independent variables):

Amount of financial reward. A third of the tasks (14 HITs, 1400 assignments) paid 5 cents, a third paid 10 cents, and a third paid 20 cents. In order to keep all my controls identical, I used the same amounts for both writing and illustration.

Provision of supplemental bonus. 6 HITs (600 assignments) mentioned a $1 bonus for the best submission, while another 6 HITs (600 assignments) promised a larger $5 bonus. The bonus was mentioned in the title of the HIT, as well as in the description which users can preview before deciding whether or not to accept a HIT. In addition, the word "bonus" was also included as a keyword for users that might search for tasks in this way. All other HITs made no mention of a potential bonus.

Stated purpose of the project. 6 HITs (600 assignments) specified - truthfully - that the purpose of the project was charity, with the proceeds from the final book being donated to a nonprofit organization that teaches digital literacy to impoverished youth in India. Conversely, another 6 HITs (600 assignments) claimed - untruthfully - that the book will be sold commercially, online and in stores, for the profit of the author. The purpose (charity or commercial) was mentioned in the title of the HIT, in the description and as a keyword. All other HITs made no mention of the purpose of this project.

Attribution of authorship. 6 HITs (600 assignments) mentioned that participants will be

credited as co-authors in the finalized book, while another 6 HITs (600 assignments) specified that participants would not be getting authorship credit when the book is published. Just like in the previous cases, the attribution of authorship (or lack thereof) was mentioned in the title of the HIT, in the description and as a keyword. All other HITs made no mention of whether participants would receive authorship credit or not.

In order to understand the interaction between factors, a total of 42 tasks were created, accounting for all possible combinations between type (written or visual), amount of reward (5, 10 or 20 cents) and the specific incentives being investigated in this study (bonus, purpose, attribution of credit). Thus, 4200 submissions were collected in total (2100 written and 2100 visual), corresponding to 42 tasks with 100 responses each.

It is also important to note that the same initial sentence was used for all writing and illustration tasks. This strategy allowed for a more accurate comparison between the conditions and facilitated the development of a universal coding scheme to assess the quality of all submissions. Importantly, it also helped avoid any differences caused by the inherent characteristics of the given narrative situation; for example, one sentence might be more narratively closed than another and therefore harder to build off of, or one narrative situation might be more difficult to draw than another.

3.3. Measures

Based on comparative analyses between these experimental conditions, the principal goal of the study was to assess the impact of the above-mentioned independent variables on the following measures (dependent variables):

Quantity of submissions. The quantity of submissions was operationalized as the amount of time (measured in hours) that it took for all 100 assignments to be completed within each task. In other words, how easy was it to reach the desired number of submissions in each experimental condition? This measure was computed by calculating the difference, in hours, between the time a task was posted and, respectively, the time that the last assignment pertaining to that task was submitted.

Quality of submissions. For the purpose of this study, the highest quality submissions were those that were most creative. The most widely accepted conceptualization of creativity has been as a combination of usefulness and novelty [26], [27]. Therefore, in assessing the quality of the submissions in this study, a coding rubric was developed to measure those two dimensions of creativity: usefulness and novelty. Both written and visual submissions were assessed for usefulness (ranging from 0 to 3 points) and novelty (also 0-3 points). To ensure maximum applicability and relevance, the coding rubric for measuring the quality of submissions was developed in an emerging fashion, after a preliminary examination of the data. After a round of refinement, a subset of 10 tasks (23.8% of the entire sample) was randomly selected for coding and intercoder reliability was assessed with the aid of ReCal2, an online software developed by Dr. Deen Freelon of American University. The results were very satisfactory and are reproduced in the tables below.

Table 1. Intercoder Reliability Results for Writing Tasks

Measure	Percent Agreement	Scott's Pi	Krippendorff's Alpha
Usefulness	99.20	0.98	0.98
Novelty	96.40	0.93	0.93

Table 2. Intercoder Reliability Results for Illustration Tasks

Measure	Percent Agreement	Scott's Pi	Krippendorff's Alpha
Usefulness	96.92	0.95	0.95
Novelty	93.28	0.90	0.90

Time spent on task. This dependent variable aimed to assess the relative effort that participants put into their work, by measuring the amount of time, in seconds, that was spent on completing each submission. This measure was computed automatically by Mechanical Turk and exported as metadata.

4. Results

The data was first cleaned up by removing compromised, incomplete and duplicate contributions. Next, the data set was analyzed in SPSS 22 via means comparison (in regards to the quantity of submissions) and ANOVAs (for usefulness, novelty, and time spent). When there was homogeneity of variances, as assessed by Levene's test for equality of variances, ANOVAs were followed by post-hoc Turkey tests to determine statistically significant contrasts between the groups. When variances were found to be not homogeneous, Welch's ANOVAs were used, followed by post-hoc Games-Howell tests.

Quantity of Submissions

Looking at the results for each experimental condition (Table 3), the hypotheses regarding bonus payments and reward amount were supported, as these factors did have a stimulating effect on the quantity of submissions. The hypotheses regarding purpose and authorship credit were contradicted, as both of these

variables had an opposite effect than had been expected: rather than encouraging potential workers to accept the task, the mention of a charitable purpose and, respectively, the possibility of gaining credit dissuaded them from participating.

Table 3. Quantity of Submissions: Summary of Results

Hypothesis	Group	# of hours till 100 submissions
H1: The **quantity of submissions** will increase when:		
a. there is a potential **bonus** payment	*control*	358
	$1 bonus	322
	$5 bonus	294
b. the **purpose** of the task is charity, and will decrease when the purpose is commercial	*control*	358
	charity	380
	comm.	328
c. **authorship credit** is offered, and will decrease when there is no credit offered	*control*	358
	credit	553
	no credit	356
d. the **amount of the reward** is greater, and will decrease when the reward is lower	*5 cents*	491
	10 cents	378
	20 cents	241

Quality of Submissions

According to the rationale outlined earlier, quality was conceptualized along two dimensions: usefulness and novelty. For the purpose of clarity, I will treat the two concepts - usefulness and novelty - separately in the following analysis.

Usefulness

In terms of the impact of bonuses, purpose, and authorship credit, although the difference in usefulness scores fits the hypotheses, the differences were not large enough to be statistically significant. Looking at the impact of the reward amounts, there was a statistically significant difference in usefulness scores between the three conditions (5 cents, 10 cents, and 20 cents): Welch's $F(2, 2435.396) = 8.937$, $p < .001$. According to the Games-Howell post-hoc test, the contrast between the lowest and the highest paid groups was statistically significant at $p < .001$, with a mean increase of .20 (95% CI [.0891, .3109]) in usefulness scores between the 5-cent condition and the 20-cent condition.

Therefore, we can conclude that only Hypothesis 2d (regarding the amount of the financial rewards) was supported to a statistically significant degree, while Hypothesis 2b (regarding the stated purpose of the project) was true for illustration tasks, but not for writing tasks. Hypotheses 2a and 2c were not supported: offering a bonus payment or providing authorship credit made little difference in terms of the usefulness of the submissions across both types of tasks.

Table 4. Usefulness of Submissions: Summary of Results

Hypothesis	Group means and standard deviations			ANOVA F value
	Group	M	SD	
H2: **Usefulness** will increase when:				
a. there is a potential **bonus** payment	*control*	1.97	1.15	
	$1	2.02	1.14	0.378
	$5	2.03	1.18	
b. the **purpose** of the task is charity, and will decrease when the purpose is commercial	*control*	1.97	1.15	
	charity	2.07	1.09	2.020
	comm.	1.93	1.22	
c. **authorship credit** is offered, and will decrease when there is no credit offered	*control*	1.97	1.15	
	credit	1.99	1.16	1.221
	no	1.89	1.19	
d. the **amount of the reward** is greater, and will decrease when the reward is lower	*5 cents*	1.89	1.20	
	10 cents	1.99	1.15	8.937**
	20 cents	2.08	1.13	

Note: ** $p < .001$

Novelty

The impact of bonus was statistically significant, $F(2, 1582) = 6.484$, $p = .002$. The Tukey test further identified a statistically significant difference between the control condition and the $5 bonus condition, which amounted to a mean increase of .196, 95% CI [.0680, .3240], $p = .001$.

In terms of the stated purpose of the project, the difference between the three conditions (control, charity and commercial) was also statistically significant, Welch's $F(2, 1047.543) = 21.159$, $p < .001$. The Games-Howell test revealed two significant contrasts (both at $p < .001$): there was a mean increase of .22 (95% CI [.1043, .3419]) between the control group and the charity group, and an even bigger increase of .33 (95% CI [.2081, .4479]) between the commercial and the charity conditions.

The difference between credit conditions (control, credit, no credit) was also statistically significant, Welch's $F(2, 1047.077) = 20.497$, $p < .001$. Novelty scores increased from the no credit group to the control group to the credit group, in that order. A Games-Howell post-hoc analysis revealed that all these increases were statistically significant: the mean increase from no credit to control (.16, 95% CI [.0459, .2686], $p = .003$) from control to credit (.15, 95% CI [.0276, .2678], $p = .011$) and from no credit to credit (.30, 95% CI [.1926, .4174], $p < .001$).

Looking at the impact of the reward amounts, the difference between the three conditions was significant, $F(2, 3671) = 8.200$, $p < .001$. According to the post-hoc Tukey test, there was a statistically significant contrast between the 5-cent and the 20-cent

groups (a mean increase of .13, 95% CI [.0492, .2097], p < .001), and between the 10-cent and 20-cent groups (a mean increase of .11, 95% CI [.0258, .1846], p=.005).

Table 5. Novelty of Submissions: Summary of Results

Hypothesis	Group means and standard deviations			ANOVA F value
	Group	M	SD	
H3: **Novelty** will increase when:				
a. there is a potential **bonus** payment	*control*	0.63	0.83	6.484*
	$1 bonus	0.74	0.89	
	$5 bonus	0.83	0.93	
b. the **purpose** of the task is charity, and will decrease when the purpose is commercial	*control*	0.63	0.83	21.159**
	charity	0.85	0.88	
	comm.	0.52	0.76	
c. **authorship credit** is offered, and will decrease when there is no credit offered	*control*	0.63	0.83	20.497**
	credit	0.77	0.84	
	no credit	0.47	0.71	
d. the **amount of the reward** is greater, and will decrease when the reward is lower	*5 cents*	0.64	0.83	8.200**
	10 cents	0.66	0.81	
	20 cents	0.77	0.88	

Note: *p < .05; ** p < .001

In conclusion, all hypotheses about novelty were supported (and with very high statistical significance levels), suggesting that the novelty of creative contributions can indeed be boosted by a bonus payment (Hypothesis 3a), a charitable purpose (Hypothesis 3b), the attribution of authorship credit (Hypothesis 3c) or a higher financial rewards (Hypothesis 3d).

Time Spent on Task

In regards to bonus payments, the difference between groups was statistically significant, Welch's F (2, 1013.039) = 7.779, p < .001. Games-Howell tests found that both increases in time spent were statistically significant: between no bonus and $1 bonus (a mean increase of 125.25, 95% CI [44.8483, 205.6539], p = .001) and between no bonus and $5 bonus (a mean increase of 85.50, 95% CI [7.2861, 163.7111], p = .028).

The differences in terms of project purpose (control vs charity vs commercial) were also significant, Welch's F (2, 972.770) = 3.484, p = .031. There was a statistically significant contrast between the commercial and charity conditions: a mean increase of 131.04, 95% CI [13.8820, 248.2046], p = .024.

In terms of credit, although the time spent by participants on the task increased from the no credit condition to the control condition to the credit condition, in that order, the difference was not statistically significant.

Finally, looking at the amount of the financial rewards, the difference between the three groups (5 cents, 10 cents and 20 cents) was significant, Welch's F (2, 2301.900) = 22.489, p < .001. The post-hoc tests revealed that all the contrasts between the conditions were statistically significant: between 5 and 10 cents (a mean increase of 54.13, 95% CI [11.9854, 96.2838], p = .007), between 10 and 20 cents (a mean increase of 129.97, 95% CI [60.7303, 199.2020], p < .001) and between 5 and 20 cents (a mean increase of 184.10, 95% CI [118.0893, 250.1121], p < .001).

Table 6. Time Spent on Task: Summary of Results

Hypothesis	Group means and standard deviations			ANOVA F value
	Group	M	SD	
H4: **Time spent on task** will increase when:				
a. there is a potential **bonus** payment	*control*	231.20	439.90	7.779**
	$1 bonus	356.45	659.21	
	$5 bonus	316.70	620.53	
b. the **purpose** of the task is charity, and will decrease when the purpose is commercial	*control*	231.20	439.90	3.484*
	charity	340.05	1075.84	
	comm.	209.01	390.25	
c. **authorship credit** is offered, and will decrease when there is no credit offered	*control*	231.20	439.90	1.621
	credit	287.26	587.85	
	no credit	244.86	547.84	
d. the **amount of the reward** is greater, and will decrease when the reward is lower	*5 cents*	202.40	378.91	22.489**
	10 cents	256.54	497.10	
	20 cents	386.50	921.44	

Note: *p < .05; ** p < .001

5. Discussion

This study makes a significant contribution towards a better understanding of creative crowdsourcing practices. In particular, the process of modifying these key factors (financial reward, bonus, purpose and authorship credit) and assessing the resulting submissions helps shape a more nuanced view of the strategies that work best when soliciting creative contributions online.

When the goal is to gather as many contributions as possible in a relatively short time span, the results of this study indicate that financial rewards work best to achieve the desired result. In other words, offering a higher reward and/or an additional bonus will lead to the timely completion of tasks, but does not always ensure the best quality and greatest effort on the part of the contributors. This conclusion is in line with existing research [8], [16]. A surprising result that contradicted the stated hypotheses was that neither a prosocial purpose nor the attribution of credit work to precipitate the completion of tasks. In particular, the attribution of credit, which had been envisioned as a significant incentive, proved to have the opposite effect, dissuading participants from accepting the tasks

in a timely manner. This effect was observed for both writing and illustration tasks (although it was significantly more pronounced for the former) and is most likely explained by the fact that people are reticent to attach their name to a project or task unless they are confident in their skills and certain that their contribution will be well received. In the case of this children's book, the fact that contributors did not have full knowledge of the final content of the book - nor its public framing and publication venues - could have made some of them less eager to contribute and be credited as co-authors on a final product that is largely outside of their control. Therefore, if authorship credit is an essential element of a collaborative creative project, sufficient details must be provided regarding the final outcome of the project, including - when applicable - legal and ethical considerations surrounding collaborative authorship.

If, however, the goal is not quantity and speed of completion, but quality - especially in terms of maximum creativity and diversity of submissions - the best strategy is to offer a bonus, emphasize the purpose of the project (if it is a charitable/prosocial one), and offer authorship credit. Indeed, for both written and visual contributions, a bonus, a charitable purpose or an attribution of authorship all promise to increase the novelty of the responses - thus increasing the overall diversity of the pool of submissions - but not necessarily their usefulness. The time spent by participants on completing the task also increases in these conditions.

Although the results in regards to the study hypotheses are very insightful, a statistical analysis cannot begin to convey the extraordinary creativity, diversity and humor that characterized the responses for both written and visual responses.

For writing tasks, most participants continued the initial sentence by having Hashtag post a "lost and found" ad for the missing shell on various websites, online community boards and social media (Snailbook, Snitter, Snailslist, etc), usually after snapping a picture of the shell with his smartphone (amusingly dubbed slimePhone, iSnail, iSlime, shellular phone, etc.). Hashtag found many friends in his adventures: characters like Trending the Slug, Underscore the Worm, Ampersand the Snake, Emoji the Turtle, Tweeter the Bird, Wifi the Walrus, Instagram the Bee, Grandmother Google, Google the Frog, Google the Goat, Google the Groundhog, Escargoogle, and many other snails that went by names like Selfie, Retweet, Backslash, Dotcom, Websurf, Asterisk, Barcode, SlowPoke, Shell Script, and Cyber. The submissions included lots of clever puns (*Looking at the empty shell, Hashtag wondered "where did Eskar go?!?")* and even rhyming (*An empty shell? What a fright! / For surely this shell is another snails's delight…/ Oh! I know what to do, I'll make it right. / I'll make it right with this tweet I write!).*

The images uploaded for the illustration tasks are perhaps even more impressive in their creativity and whimsy, and in the tremendous effort that participants evidently put into their work. There was a wide range of visual styles represented, as well as a multitude of visual media, including digital illustrations, 3D renderings, hand drawing, acrylics, watercolors, collages, and found objects.

An optional question at the end of each task asked participants for open-ended feedback or comments about the task or about Hashtag's story, and was meant as a space for contributors to voice their opinions and provide unrestricted input. Surprisingly, this question elicited an impressive number of responses, with approximately half of the participants choosing to fill it in. Most responses expressed the participants' appreciation of the HIT; in fact, many workers just used this space to thank me for an enjoyable task. Some mentioned the effect that completing the HIT had on them, often in ebullient terms ("Feeling very relaxed after this joyful thing!"; "I was in a bad mood, and drawing Mr. Hashtag cheered me up!"). Participants also liked the idea of collectively writing and illustrating a story, one sentence at a time, and some stated that they wanted to try out this idea with their kids or students ("I'm a preschool teacher and love to create and teach kids stories and crafts. This HIT made me think to do some similar sort of project with my kids").

Only a minority of respondents used the feedback space to provide constructive criticism about the HITs. Of these, the most common observations concerned the need to pay more for illustration tasks, to let them write as much as they want for the writing HITs (this was implied in my instructions which only stipulated a minimum length, i.e. a sentence, but perhaps that was not entirely clear) and to provide more information about the visual characterization of Hashtag the Snail. On this last point, respondents advised that "you should have an illustration of your character, so any future artist contributing to your project have a reference to work from." Another, who identified as a "graphic designer / illustrator" agreed: "For something like this the illustrator will need as much data as you can possibly provide about the character you have in mind. Personality is one of the most important traits in order to design Hashtag."

The amount of the financial reward - whether the task paid 5, 10 or 20 cents - had little impact on the quality of the written submissions and the time spend by the participants on the task. For illustration tasks, on the other hand, quality and effort were indeed higher

when the reward was greater, presumably because the pay for illustration tasks was deemed to be too low, given the amount of time it took to complete these drawings and the comparative difficulty of this task versus the written tasks. The decision to offer the same rewards (5, 10 and 20 cents) for both writing and illustration tasks was in order to keep all controls identical and avoid skewing the results. However, the feedback from participants in the post-task optional question made it clear that they considered the pay for illustration HITs to be insufficient for the time they devoted to the task. The relatively low reward for illustration tasks also resulted in many rebellious contributions and rampant cheating, which did not occur for writing HITs. Participants submitted duplicate images from different accounts, uploaded images of snails from the Internet and - even more interestingly - submitted a wide variety of rebellious contributions. The content of the rebellious contributions was nonetheless very amusing, if oftentimes perplexing. Most rebellious contributions were pictures of animals and babies (including a couple of babies with snail shells on their backs, taken from the Internet). Others were personal pictures, screenshots and even a couple of comics. In one of the more interesting cases, someone uploaded a 3D digital model of a house that they had created, noting in the feedback space that they could not draw a snail but they are very talented at drawing buildings if I am ever in need of those skills.

6. Limitations

This study also presents several limitations that must be acknowledged. Firstly, the fact that a series of related tasks were posted on the same platform - albeit spaced out over a period of time - meant that participants could have seen and even participated in multiple tasks. This is problematic for two reasons: one, if participants saw multiple related tasks, they could have realized that certain key variables were being modified; and two, if workers participated in more than one conditions, that violates the assumption of independence of samples. Given the available options when creating and posting tasks on Mechanical Turk, there is no easy way to avoid this challenge; however, this is a significant challenge that merits further discussion and investigation. A recent study by Chandler, Mueller and Paolacci has shown that researchers using Mechanical Turk are largely unaware of the possibility that workers might participate in related experiments [28]. The authors caution that, although the Mechanical Turk worker pool can seem almost inexhaustible (especially in comparison to participant pools used in traditional research studies), duplicate workers are more common than researchers assume. Beyond post-hoc data cleaning - which is very common but sometimes problematic - the authors suggest a few strategies that researchers can rely on to avoid this problem. A simple solution would be running multiple related experiments through one single link within the same HIT, but this is not always feasible - as in the case of the present study - because such a strategy does not allow the researchers to modify key variables like reward amounts and task details. Other potential strategies, depending on a study's research design, are to assign Qualifications to workers who are prescreened (within the Mechanical Turk platform), or, for those researchers with significant programming experience, to use the Mechanical Turk API (Application Program Interface) in order to modify the HIT parameters and exclude certain workers [28]. Finally, and also depending on the research design, an alternative option would be to use an external research platform, such as Qualtrics, in order to prescreen Mechanical Turk workers or to randomly assign them to different experimental conditions within the same study.

Self-selection bias is another potential limitation in this case, as it could be that workers with certain qualities are attracted to HITs with certain advertised incentives. Thus, it could be the individual, and not the incentive, which influences participation. As Rogstadius and colleagues note, this is a particularly challenging issue for studies on motivation, "as self- selection is an inherent aspect of a task market" [15]. Furthermore, research has shown that workers frequently talk about requesters and share information about tasks [29], which is also problematic for such studies, because it affects task selection, as well as letting workers know about related studies.

Another limitation has to do with the accuracy of the time spent variable. This variable was computed automatically by Mechanical Turk based on the time elapsed between task acceptance and submission. However, there is no guarantee that the users spent all that time actively working on the task: they could have been multitasking or could have even stepped away from their computer.

Finally, another important limitation pertains to the generalizability of the findings beyond the Mechanical Turk population. In addition to the demographical particularities of Mechanical Turk workers, there are other important differences to take into account when comparing Mechanical Turk samples to traditional research samples. In a 2013 study, Goodman, Cryder & Cheema found that turkers are less likely to pay attention to experimental materials, and more likely to rely on the Internet to find answers, even when there is no incentive to submit a

correct response [30]. Both of these considerations are relevant to the current study: the former, as the authors observe, can reduce the statistical power of the experimental research, while the latter can help explain why so many workers cheated on the illustration task by uploading pictures from the Internet. Finally, studies show that Mechanical Turk participants also have idiosyncratic attitudes about money, that are not representative to those of a normal population but are in fact similar to the attitudes of student populations [29]. This is a very interesting observation, which could play a significant role in terms of the impact that financial rewards had on the present study's dependent variables.

7. Conclusion

Attaining a better understanding of the incentive structures driving online participation has been an important goal of recent crowdsourcing scholarship, and the present research makes a significant contribution to this body of work, specifically in regards to creative crowdsourcing processes. The results of this study shed light on the conditions that most effectively foster creative participation online (as well as those that fail to do so), investigating both the inputs and the outputs of creative participation. In view of the versatility and growing popularity of creative crowdsourcing projects, these findings will hopefully be useful not only to scholars, but also to companies, artists and practitioners who would like to rely on open public participation for their creative projects.

The relative importance of a prosocial purpose and the attribution of authorship credit are, in particular, novel findings worthy of deeper consideration in future research. In terms of the larger purpose of participatory projects, more research is needed on the impact of task meaningfulness on the quality of participation. Interestingly, the relationship between purpose and quality observed in this study - specifically, that the quality of the submissions was higher when the stated purpose of the project was charitable - stands in contrast to the findings of Chandler and Kapelner, who concluded, also based on a Mechanical Turk study, that the framing of a task as meaningful (in this case, labeling tumor cells for cancer research) does not boost the quality of the resulting submissions [7]. There is a significant gap in the literature as to the role of authorship credit in crowdsourcing projects - creative or otherwise - so this is a key area where future scholarship is needed. The present study found that attributions of authorship deepen participants' investment in the context of creative projects, but it is uncertain whether the same conclusion would hold true in other contexts. In the same time, it is also important to note that, as illustrated by the rise of crowdsourced art and participatory cultures, notions of authorship are in flux; consequently, there is a need to account for new forms of authorship, especially ones that are quintessentially collective or distributed.

References

[1] Howe, J. (2006). The rise of crowdsourcing. *Wired*, 14(6), Retrieved from http://www.wired.com/wired/archive/14.06/crowds.html

[2] Literat, I. (2012). The work of art in the age of mediated participation: Crowdsourced art and collective creativity. *International Journal of Communication* 6: 2962-2984.

[3] eYeka (2015). The state of crowdsourcing in 2015. Research report. Retrieved from https://en.eyeka.com/resources/reports#CSreport2015

[4] Bayus, B. L. (2013). Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. *Management science*, *59*(1), 226-244.

[5] Keating, M., Rhodes, B., & Richards, A. (2013). Crowdsourcing: a flexible method for innovation, data collection, and analysis in social science research. *Social media, sociality, and survey research*, 179-201.

[6] Brabham, D. C. (2010). Moving the crowd at Threadless. *Information, Communication & Society*, *13* (8), 1122-1145.

[7] Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, *90*, 123-133.

[8] Zheng, H., Li, D., & Hou, W. (2011). Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce 15* (4), 57–88.

[9] Simula, H. (2013, January). The rise and fall of crowdsourcing?. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 2783-2791). IEEE.

[10] Ryan, R. M., & Deci, E. L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist*, *55*(1), 68–78.

[11] Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, *25*(1), 54– 67.

[12] Deci, E. (1975). *Intrinsic motivation.* New York, NY: Plenum Press.

[13] Deci, E., Koestner, R. & Ryan, R. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125* (6): 627–668.

[14] Frey, B. (1997). *Not just for the money: An economic theory of personal motivation*. Cheltenham, U.K.: Edward Elgar.

[15] Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., & Vukovic, M. (2011). An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

[16] Mason, W. A., Watts, D. J. 2009. Financial incentives and the performance of crowds. In Proc. ACM SIGKDD Workshop on Human Computation, pp. 77-85. ACM Press.

[17] Walter, T., & Back, A. (2011). Towards measuring crowdsourcing success: An empirical study on effects of external factors in online idea contest.

[18] Malone, T. W., Laubacher, R. & Dellarocas, C. (2010). The collective intelligence genome. Sloan Management Review, 51:3, 21-31.

[19] Antikainen, M.J. and Väätäjä, H.K. (2010) 'Rewarding in open innovation communities – how to motivate members', *Int. J. Entrepreneurship and Innovation Management*, Vol. 11, No. 4, pp.440–456.

[20] Shaw, A. D., Horton, J. J., & Chen, D. L. (2011, March). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 275-284). ACM.

[21] Kaufmann, N., Veit, D., and Schulze, T. (2011). More than fun and money: Worker motivation in crowdsourcing a study on Mechanical Turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems*, Detroit, MI.

[22] Ipeirotis, P. (2010) The new demographics of Mechanical Turk. Retrieved from http://behind-the-enemy-lines. blogspot.com/2010/03/ new-demographics-of-mechanical-turk.html

[23] Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, 6(1):3 –5, Jan. 2011.

[24] Marvit, M.Z. (2014, February 4). How crowd workers became the ghosts in the digital machine. *The Nation*. Retrieved from http://www.thenation.com/article/178241/how-crowdworkers-became-ghosts-digital-machine#

[25] Ipeirotis, P. (2010). Demographics of Mechanical Turk. NYU Center for Digital Economy Research Working Paper CeDER-10-01. Retrieved from http://www.ipeirotis.com/wp-content/uploads/2012/02/CeDER-10-01.pdf

[26] Amabile, T.M. (1983). The social-psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology, 45*(2), 357-376.

[27] Amabile, T.M. & Pillemer., J. (2012). Perspectives on the social psychology of creativity. *Journal of Creative Behavior, 46*(1), 3-15.

[28] Chandler, J., Mueller, P. & Paolacci, G. (2014). Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers. *Behavior Research Methods, 46* (1), 112-130.

[29] Yin, M., Gray, M. L., Suri, S., & Vaughan, J. W. (2016,). The Communication Network Within the Crowd. In Proceedings of the 25th International Conference on World Wide Web (pp. 1293-1303). International World Wide Web Conferences Steering Committee.

[30] Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*, 213–224.